2

GL-TR-90-0103

Visibility Data Filters for Europe

Bret A. Schichtel and Rudolf B. Husar

Center for Air Pollution Impact and Trend Analysis
Washington University
St. Louis, MO 63130

14 April 1990

Scientific Report #2

Approved for public release; distribution unlimited

GEOPHYSICS LABORATORY
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
HANSCOM AIR FORCE BASE, MA 01731-5000

"This technical report has been reviewed and approved for publication"

Frank W. Gibson
(Signature)
Frank W. Gibson
Contract Manager

Robert R. [illegible]
(Signature) for
Donald E. Bedo, Chief
Atmospheric Optics Branch

FOR THE COMMANDER

R Earl Good
(Signature)
R. Earl Good, Director
Optical and Infrared Technology Division

This report has been reviewed by the ESD Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS).

Qualified requestors may obtain additional copies from the Defense Technical Information Center. All others should apply to the National Technical Information Service.

If your address has changed, or if you wish to be removed from the mailing list, or if the addressee is no longer employed by your organization, please notify GL/IMA, Hanscom AFB, MA 01731. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document requires that it be turned.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release;
distribution unlimited

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
Washington University

5. MONITORING ORGANIZATION REPORT NUMBER(S)
GL-TR-90-0103

6a. NAME OF PERFORMING ORGANIZATION
Washington University

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION
Geophysics Laboratory

6c. ADDRESS *(City, State, and ZIP Code)*
Center for Air Pollution Impact & Trend Analys
Campus Box 1124
St. Louis, MO 63130

7b. ADDRESS *(City, State, and ZIP Code)*
Hanscom Air Force Base, MA 01731-5000

8a. NAME OF FUNDING / SPONSORING ORGANIZATION

8b. OFFICE SYMBOL *(If applicable)*

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
F19628-87-K-0003

8c. ADDRESS *(City, State, and ZIP Code)*

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
62101F	7670	15	AO

11. TITLE *(Include Security Classification)*
Visibility Data Filters For Europe

12. PERSONAL AUTHOR(S)
Bret A.Schichtel and Rudolf B. Husar

13a. TYPE OF REPORT
Scientific 2

13b. TIME COVERED
FROM 89/09/19 TO 90/04/14

14. DATE OF REPORT *(Year, Month, Day)*
90/04/14

15. PAGE COUNT
44

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Visibility, European visibility

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

The purpose of the report is to present the methodology for filtering the European meteorological visibility data from undesirable and erroneous data. Seven data filters were devised and imposed on the European synoptic visibility data set. The data set consisted of fourteen years of meteorological data (1973-1986) for about 1600 station in Europe. The European data set was extracted from the DATSAV global weather database maintained by the U.S. Air Force, ETAC, Scott Air Force Base.[1] The raw meteorological data set consisted of over 1000 magnetic tapes containing about 30 gigabytes of data. The first step in the data processing involved compacting the data set into a binary form, which reduced the data size to a 3 gigabytes. Next, from the daily

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED □ SAME AS RPT. □ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Frank Gibson

22b. TELEPHONE *(Include Area Code)*
(617) 377 3665

22c. OFFICE SYMBOL
GL/OPA

83 APR edition may be used until exhausted.
All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
unclassified

U.S. Government Printing Office: 1985-507-047

visibility data, the quarterly cumulative distribution functions for extinction coefficient B_{ext} and visibility were computed. Most of the subsequent data filtering was performed using the distribution functions and the Voyager data exploration software.

The data cleaning filters fell into three categories: 1. precipitation and humidity filters, 2. filters based on the year to year fluctuation of extinction coefficient, and 3. filters that eliminated entire stations. The precipitation and humidity filter was imposed on the hourly data, while the remaining filters operated on the distribution functions. The main cause of poor data was identified to be the visibility threshold, that is the maximum distance reported for a given station. The visibility threshold was detected using the shape and time trend of B_{ext} quantiles. A visibility threshold truncates the distribution function and also causes the lower percentiles to be invariant with time. It was determined that the 75th percentile of B_{ext} is a robust measure of the extinction coefficient, relatively uninfluenced by the visibility threshold. The resulting data base is suitable for input to radiative transmission and transfer models, global climate models, air pollution studies as well as to global biogeochemical explorations.

ACKNOWLEDGEMENTS

The processing of the DATSAV database was a three year effort during which many individuals contributed. The initial data processing was conducted by Feliks Golenko. Some of the data processing routines were prepared by Masuri Hashim. The development of the visibility data filters was conducted by Christer Morales of Stockholm University. His insight and help is gratefully acknowledged.

1. INTRODUCTION

This report presents the visibility climate for Europe based on data collected at over 1700 meteorological stations. Its main purpose is to discuss the data quality, the data filtering procedures, and present the quality controlled visibility climate data for Europe.

The primary utility of this aerosol climatology is that it provides suitable aerosol input data to atmospheric radiative transmission models such as LOWTRAN and FASTCODE[2]. It is believed that beyond these immediate applications this aerosol database will find application in global climate models, air pollution studies as well as global biogeochemical cycle studies.

1.1 Related Reports

This report is one of the summary reports presenting the results from the past three years of research as part of this contract. Other summary reports include the *"Organization, Access, and Exploration Facilities for Large Geophysical Databases"*, dated September 22, 1989. It describes the data organization principles applied to the visibility and aerosol databases. Additional summary reports will include presentation of: Visibility Climate for North America and Asia; Characterizational Aerosol Types; Interfacing of Aerosol Climate Data with LOWTRAN and FASTCODE Models; Database and Software User Manual.

1.2 Raw Data Source and Characteristics

The data set for this phase of the study consisted of fourteen years of meteorological data (1973-1986) for about 1600 station in Europe. This data set was extracted from the DATSAV global weather database maintained by the U.S. Air Force, ETAC, Scott Air Force Base, IL.

The raw meteorological data set consisted of over 1000 magnetic tapes containing about 30 gigabytes of data. The first step in the data processing involved compacting the data set into a binary form, which reduced the data size to a more manageable 3 gigabytes. Next, from the daily visibility data, the cumulative distribution functions for extinction coefficient were computed. Most of the subsequent data filtering was performed using the aggregated distribution functions and the Voyager data exploration software. A detailed description of the data pre-processing steps is beyond the scope of this report. It suffice to state that most of the time and effort was invested in reading and processing of 30 gigabytes of meteorological data.

2. DATA DISTRIBUTION FUNCTIONS

Visibility is a measure of atmospheric optics, and it is inversely proportional to the atmospheric aerosol concentration. The visibility is often converted to the extinction coefficient by the Koschmieder relationship, B_{ext} = 3.9/Visibility (km^{-1}) which is a direct measure of haze. The Koschmieder constant of 3.9 corresponds to the visibility threshold of 2% contrast. It has been shown by various studies (e.g. Middleton, 1952[3]) that for atmospheric observations a 5% contrast threshold is more appropriate with the corresponding Koschmieder constant of 3.0. In the following analysis we have used 3.0 to convert the visual range values to extinction coefficient.

2.1 Raw Visibility Data

A problem with visual measurements is that a distance limit usually exists beyond which the visual range is not resolved. This is due to either a lack of markers, or to observation rules that do not require reporting visibility beyond this limit. This limit causes threshold values to appear in the data as illustrated in Figure 1 for Frankfurt, FR Germany. The threshold is manifested by the straight line at 11000 m visibility after 1968. Evidently, the threshold varied significantly prior to 1968. The consequence of the threshold value is that averaged data are biased by the artificial threshold. The fraction of the data that occurs at the visibility threshold have to be discarded, because it is not known what is the actual value for those data points.

Most of the effort in filtering biased data pertains to the identification and discarding of values at the visibility threshold. Our procedures for such biased data identification utilized the construction of cumulative distribution functions for visibility and extinction coefficient. In this study the data were grouped into six cumulative percentiles, the 10th, 25th, 50th, 75th, 90th, and 95th. Separate percentiles were calculated for each station, quarter, and year.

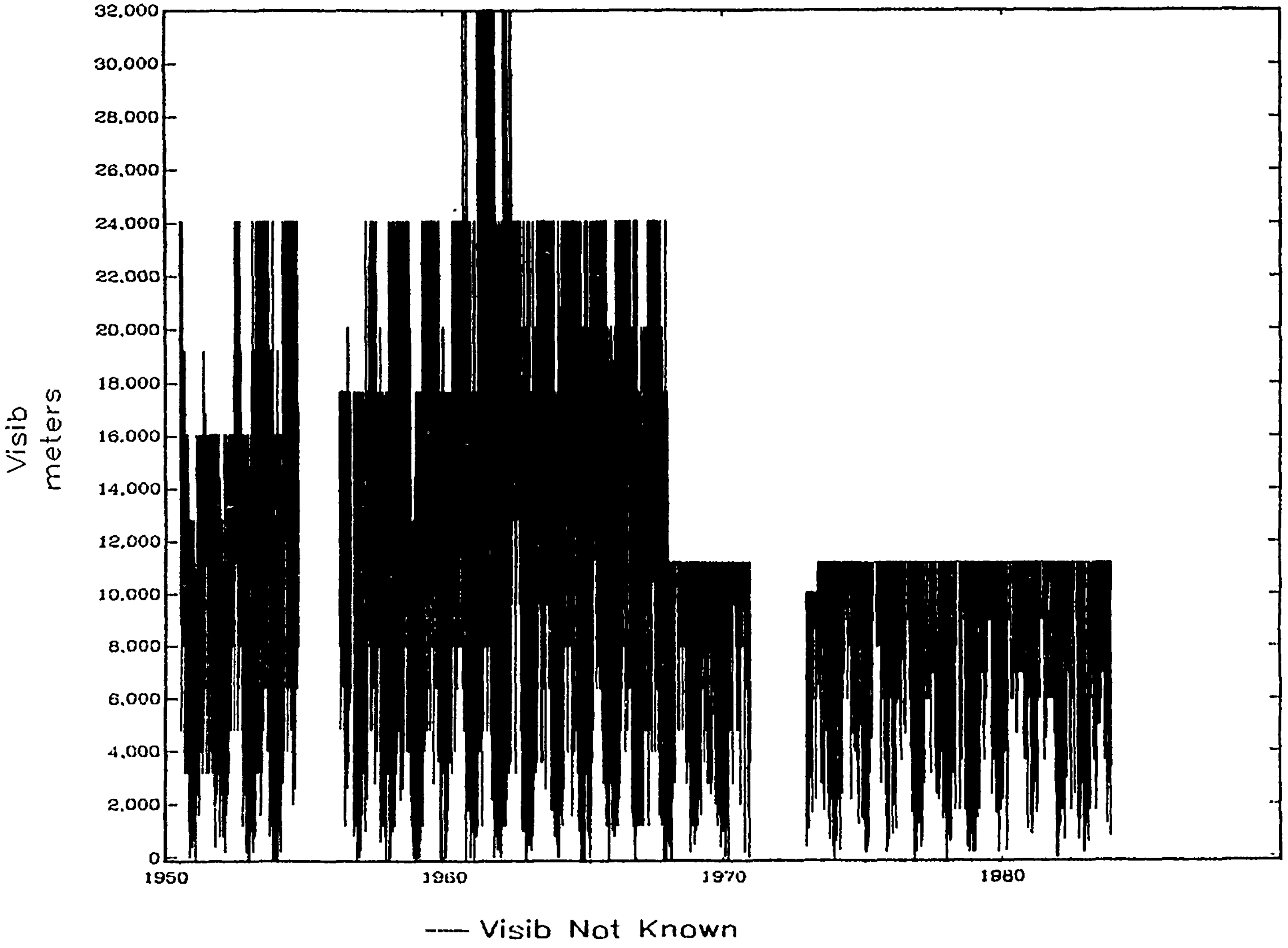

Figure 1. Trend of visual range for Frankfurt, FRG. Note the threshold visibility at 11000 m after 1968. Prior to 1968, the threshold was variable.

The threshold visual range tends to influence the higher percentiles of visibility while the distribution function for lower visibilities is unbiased by the threshold. The value of the threshold visibility changes significantly from one station to another. Furthermore, for a given station this threshold can change from one time period to another as illustrated in Figure 1. The technique for threshold filtering needs to be sufficiently sensitive to incorporate these effects.

2.2 Log-Normal Distribution

In previous research visibility data have been shown to fit a log-normal distribution (Husar et al 1979[4]; Malm et al 1981[5]). This is shown to be also true in Europe as illustrated in Figure 2. Table 1 contains the filtered and raw B_{ext} for selected stations. The filtered B_{ext} represents the extinction coefficient passed through the quality control filters as described in section 3. The raw B_{ext} represents the B_{ext} percentiles calculated from raw visibility data. Inspection of other seasons and stations confirmed the validity of the log-normal assumption. The fit to the log-normal distribution is particularly good between the 25 and 75 percentiles. Deviations from log-normality are evident at the extremes of the distribution function (e.g. Vicenza station).

Table 1. The data used to make cumulative distribution plots.

Stations	Percentiles						
	5	10	25	50	75	90	95
GUTERSLOH, GERMANY							
Bext	–	164	242	406	661	1022	1308
Converted Bext	107	150	240	416	685	1173	–
VICENZA, N. ITALY							
Bext	–	246	355	706	1350	2110	2510
Converted Bext	168	222	333	695	1570	3070	–
BURRINGTON, ENGLAND							
Bext	–	130	185	315	550	890	1160
Converted Bext	116	128	190	350	680	1335	–
BIARRITZ/ANGLET, FRANCE							
Bext	–	204	240	252	353	530	648
Converted Bext	112	168	210	250	330	490	–
CABO CARVOEIRO, SPAIN							
Bext	–	149	149	149	149	190	245
Converted Bext	149	149	149	149	149	185	–

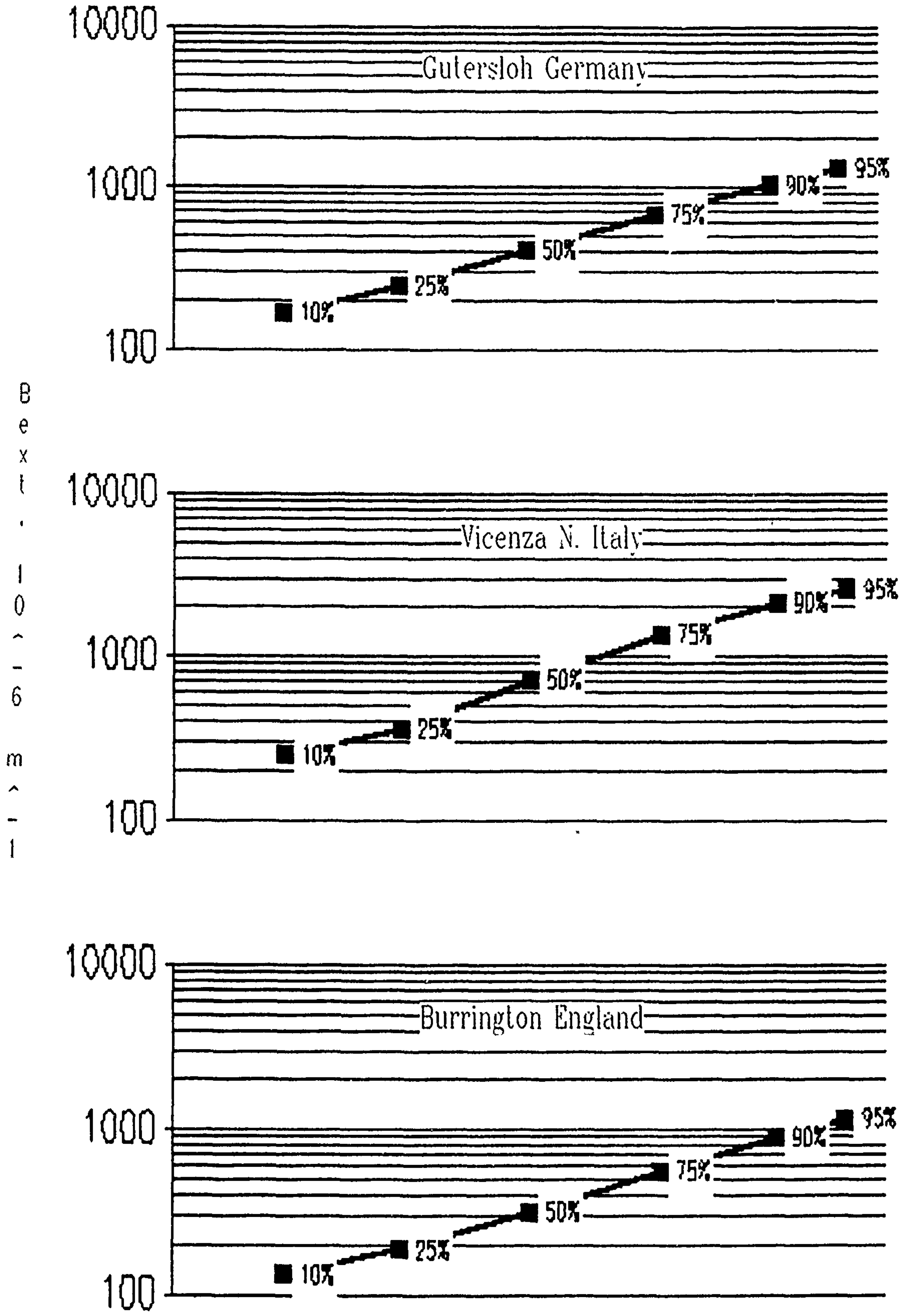

Figure 2. Cumulative frequency distribution functions for B_{ext} at three stations with good data.

A log-normal distribution only requires two parameters to give a full description of the data. Essentially, it is described by the mean logarithmic standard deviation. The two parameters used for this study are the 75th percentile, and the logarithmic standard deviation, σ_g. The 75th percentile was chosen, because for most stations it was uninfluenced by a threshold value. Given the 75th percentile and σ_g other percentiles and statistical parameters such as the mean can be found.

2.3 Stations With Incomplete Distributions

The above examples utilized data from stations where the threshold visual range was high and did not influence the distribution functions significantly. However, for most stations, the threshold values occur at lower visual range, such that the distribution functions are distorted (truncated). For such stations the log-probability plot of the visual range does not conform to the log-normal distribution, particularly at high visual ranges. Consistently, the log-probability plots deviate from a straight line.

Since the complete distribution function data can be fitted by a log-normal distribution, the percentiles below the threshold value can be estimated through extrapolation of valid data. The cumulative distribution for Biarritz/Anglet, France is shown in Figure 3a. As can be seen the percentiles above 50% form a straight line, but deviates at percentiles below 50% where the data are influenced by the threshold. If the line is extended (the doted line on the graph) the lower percentiles can be estimated. In Figure 3b Cabo Carvoeiro Spain, the cumulative distribution shows that only the 75, 90, and 95 percentiles are above the threshold. For this station 75% of the data are at the threshold limit, and hence they have the same extinction coefficient. This is the reason that the percentiles blow 75% are all at the same value of B_{ext}. This flatness of the cumulative distribution curve is a signature for truncated data. In fact, this signature is used to filter out data influenced by the threshold.

Some stations also exist where more than 75% of the data is influenced by the threshold. For these stations a reliable extrapolation of the data can not be made. Such stations were the targets of the filters described in section 3.

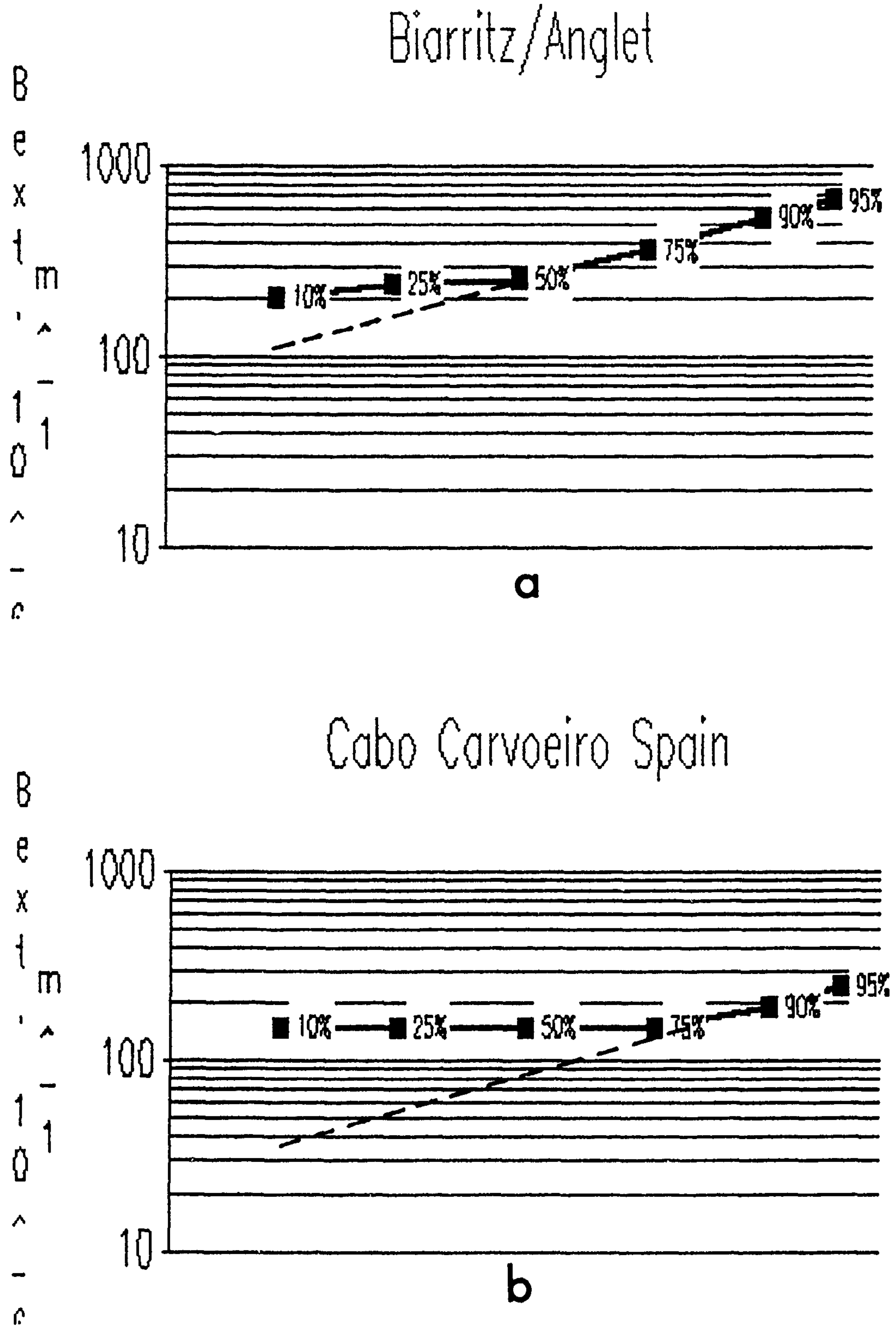

Figure 3. Cumulative frequency distribution of B_{ext} for an adequate (Biarritz/Anglet) and a poor (Cabo Carvoeiro) station.

2.4 Time Plots of the Percentiles

The time series of percentiles can also be used to detect anomalies and systematic errors in the data. Figure 4 contains plots of the 25th, 50th, 75th, and 90th percentiles for three of the stations used to make the cumulative distribution plots (Figures 2 and 3).

Figure 4a, is the time plot for Gutersloh Germany, FRG. This figure shows that the 50th, 75th, and 90th percentiles are significantly separated for each year. Furthermore, the year to year variation of a given percentile ranges between 10 to 20 percent of the mean. Such variations are considered normal noise for a meteorological variable. The 25th percentile showed much less variation and it is likely that it is under the influence of the visibility threshold. Because of the above characteristics Gutersloth is considered to be a good station.

The time plot for Biarritz/Anglet, France is represented in Figure 4b. The trend graph shows that the 25th and 50th percentiles have the same values over most of the time period. This is a signature that these percentiles are influenced by the threshold value. The 75th and 90th percentiles, on the other hand, are well separated from each other and show year to year variation. This station is considered acceptable. This graph clearly illustrates the benefit of using the 75th percentile.

Observing the trends of percentiles for Biarritz/Anglet, France from 1973 through 1986 shows a decline of the 25th and 50th percentile in 1984. On the other hand the 75th and 90th percentile show a slight increase. This inconsistency is due to the fact the 25th and 50th percentiles show changes due to threshold limit changes, while the 75th and 90th percentiles represent the real trends due to atmospheric effects.

Figure 4c, is a time plot of Cabo Carvoeiro, Spain. This figure shows that the 25th, 50th, and 75th percentiles all lie at the same value for the entire time period. The 90th percentile also lies at this value except for few years. As for Biarritz/Anglet, France this is a sign of the influence of the threshold value. Because all of the percentiles are influenced by the threshold value this is considered to be a poor station.

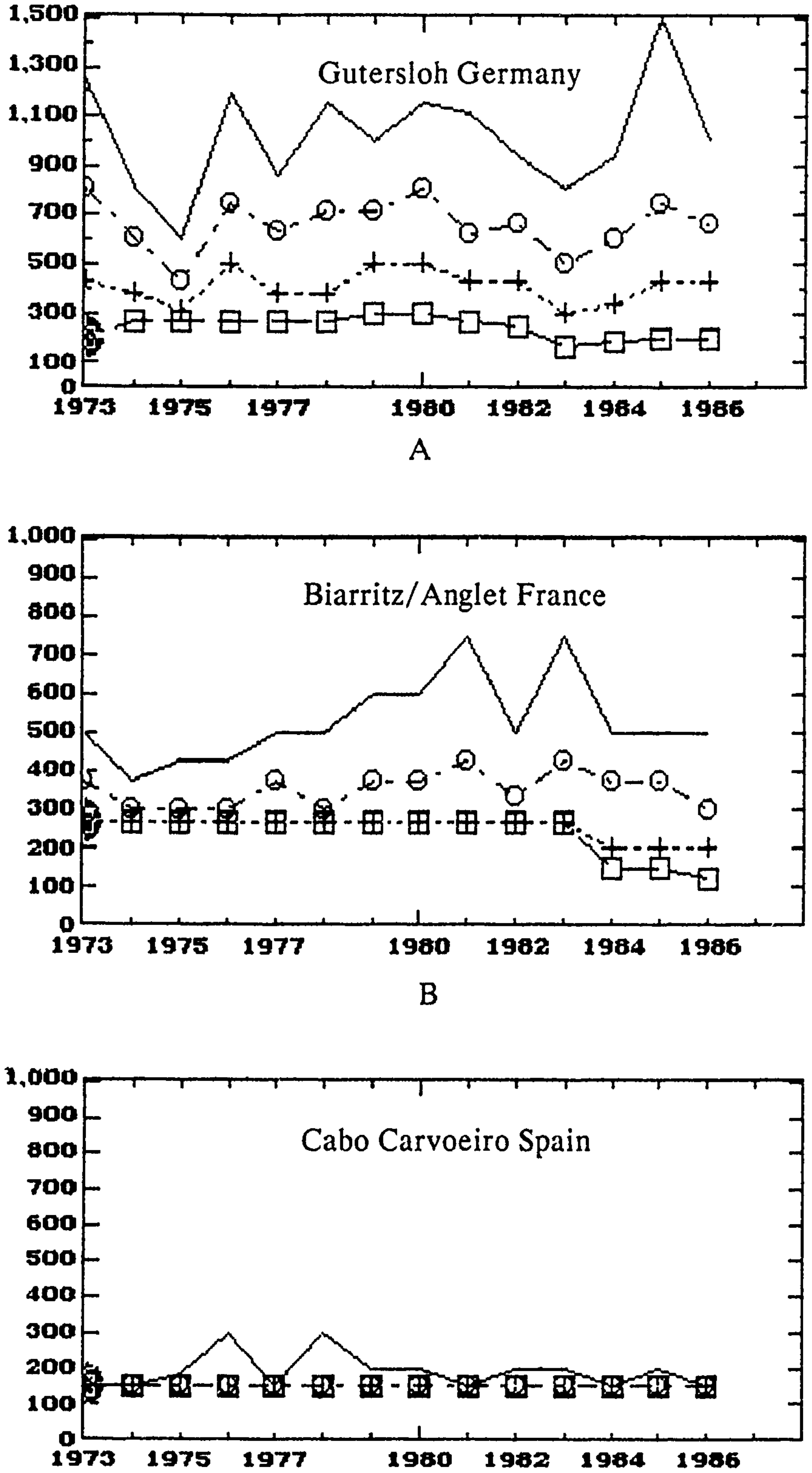

Figure 4. Time trend of B_{ext} percentiles for a good (a), satisfactory (b) and a poor (c) station.

3. DATA QUALITY FILTERS

Deriving aerosol climatological information from surface visual range observations requires significant modifications to the raw data. These modifications involves elimination of bad data arising from poor observations and recordings. Other data are eliminated due to the fact that they are strongly influenced by weather factors other than aerosols. These include precipitation, high humidity, very low clouds, and other natural phenomena.

In order to identify and eliminate these undesirable influences on the aerosol data, seven different data quality filters were designed and applied. The design of these filters was guided largely by intuition. The development of these data filters also follows the work reported by Morales et al. (1986)[6]. The design of these filters was also aided by the data exploration software Voyager by Lantern Corporation.[7] The information science basis of the Voyager software is described by our previous report: *"Organization, Access, and Exploration Facilities for Large Geophysical Databases."*

This section describes the rationale and implementation of the seven data quality filters. It also shows the consequences of imposing these filters. Finally, the clean data are presented showing the aerosol extinction data after all data filters were applied. The order of implementation of the filters influences their effects. The sequence in which the filters are discussed represents the order they were applied to the data.

3.1 Precipitation Filter and Humidity (1)

The purpose of this filter is to remove data that occurs during precipitation, high humidity, or low cloud cover. Detecting precipitation events is best done by examining the precipitation flags in the synoptic observation data. These are recorded under code WW1, WW2, etc. Unfortunately for most of the European stations contained in the DATSAV database, the precipitation flags are missing. For this reason, we constructed the precipitation and humidity filter using variables that are available in the database: dewpoint depression, visibility and ceiling. The conditions for an acceptable value are stated below:

Dew Point Depression	$< 1^{\circ}$K
Visibility	< 1000 meters
Ceiling	< 120 meters

The dewpoint depression of less than one degree K corresponds to relative humidity greater than 95%. This condition is imposed on the grounds that hygroscopic aerosols grow significantly at humidities greater than 95%. At these humidities aerosols tend to grow into hygrometers such as fog, snow, or precipitation.

Figure 5 shows the time plot of the raw visibility data, dewpoint depression, and cloud ceiling for one month in Frankfurt, FRG. The influence of the dewpoint depression on the visibility can be seen in this figure. The periods of low dewpoint depression have corresponding period of low visibility. All data points with dewpoint depression below the line drawn were discarded.

Visibility of less than one kilometer is not attributable explicitly to aerosols. In most cases the visibilities below 1000 meters occur due to precipitation or other high humidity events. Notable exceptions are fires, sand storms, or extreme pollution episodes. In this aerosol climatology, we have assumed that low visibility events (< 1000 m) occur only due to fog and precipitation. Consistently this condition generally represents a precipitation filter.

The condition, ceiling < 120 meters detects and eliminates low cloud events when the cloud height is < 120 meters above ground. Here, it is assumed that under these conditions the surface visibility conditions are significantly effected by the low clouds.

Many of the above three conditions occur simultaneously. See Figure 5. However, there are instances where only one of these condition occur. In other cases significant data may be missing such as dewpoint depression. In that case the ceiling or visibility condition is activated to reject an undesirable data point.

The precipitation filter is applied to the hourly data. The reason for this is that this filter required multiple meteorological variables as inputs. The remaining filters operate on the distribution functions.

The effect of the precipitation filter on the B_{ext} can be determined from the contour maps in Figure 6, 7, and 8. Appendix A contains a description of the process used to create these spatial contour maps. Figure 6 is a map of the extrapolated extinction coefficient from each observation site for the 75th percentile. Figure 7

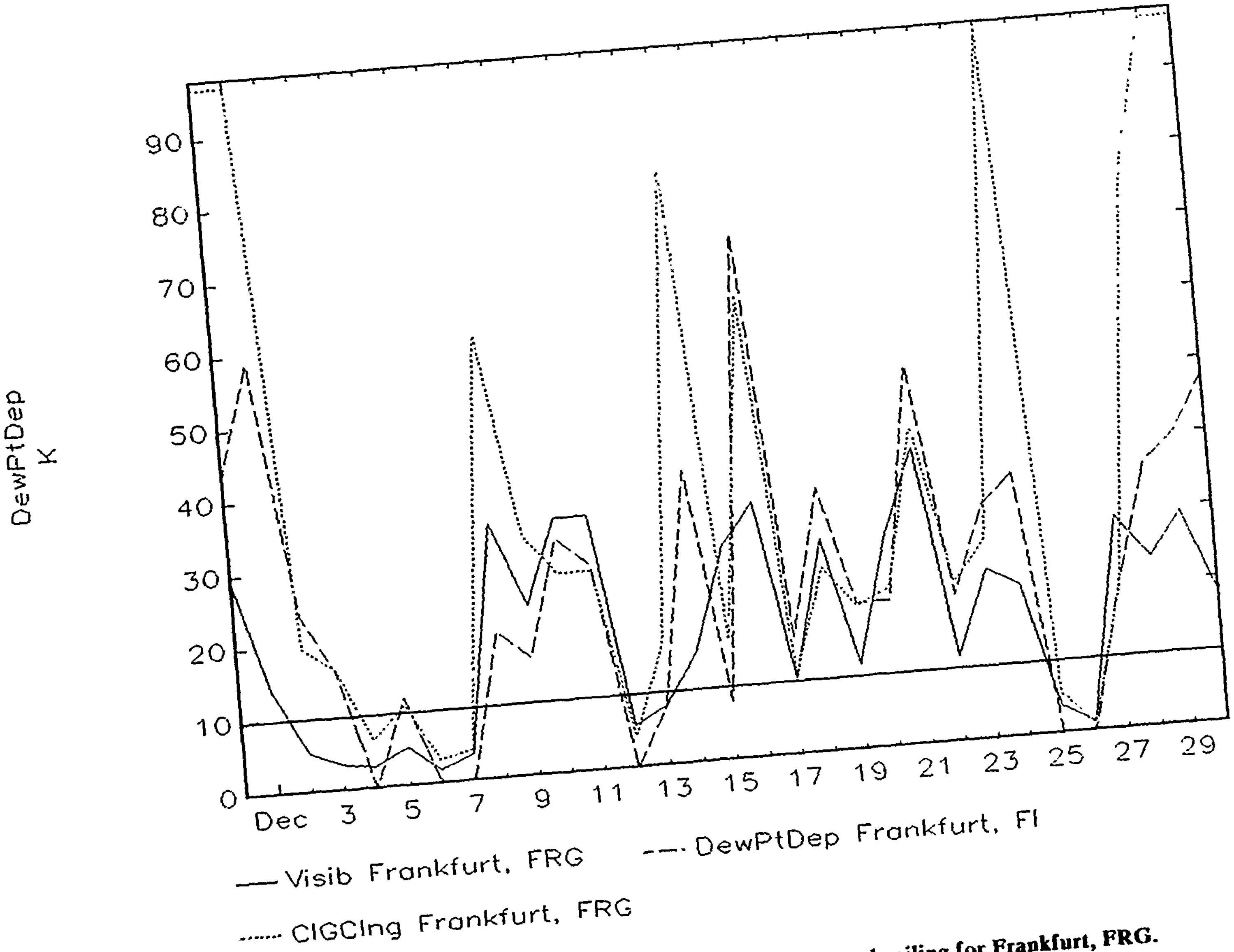

Figure 5. Time series of visibility, dew point depression, and ceiling for Frankfurt, FRG.

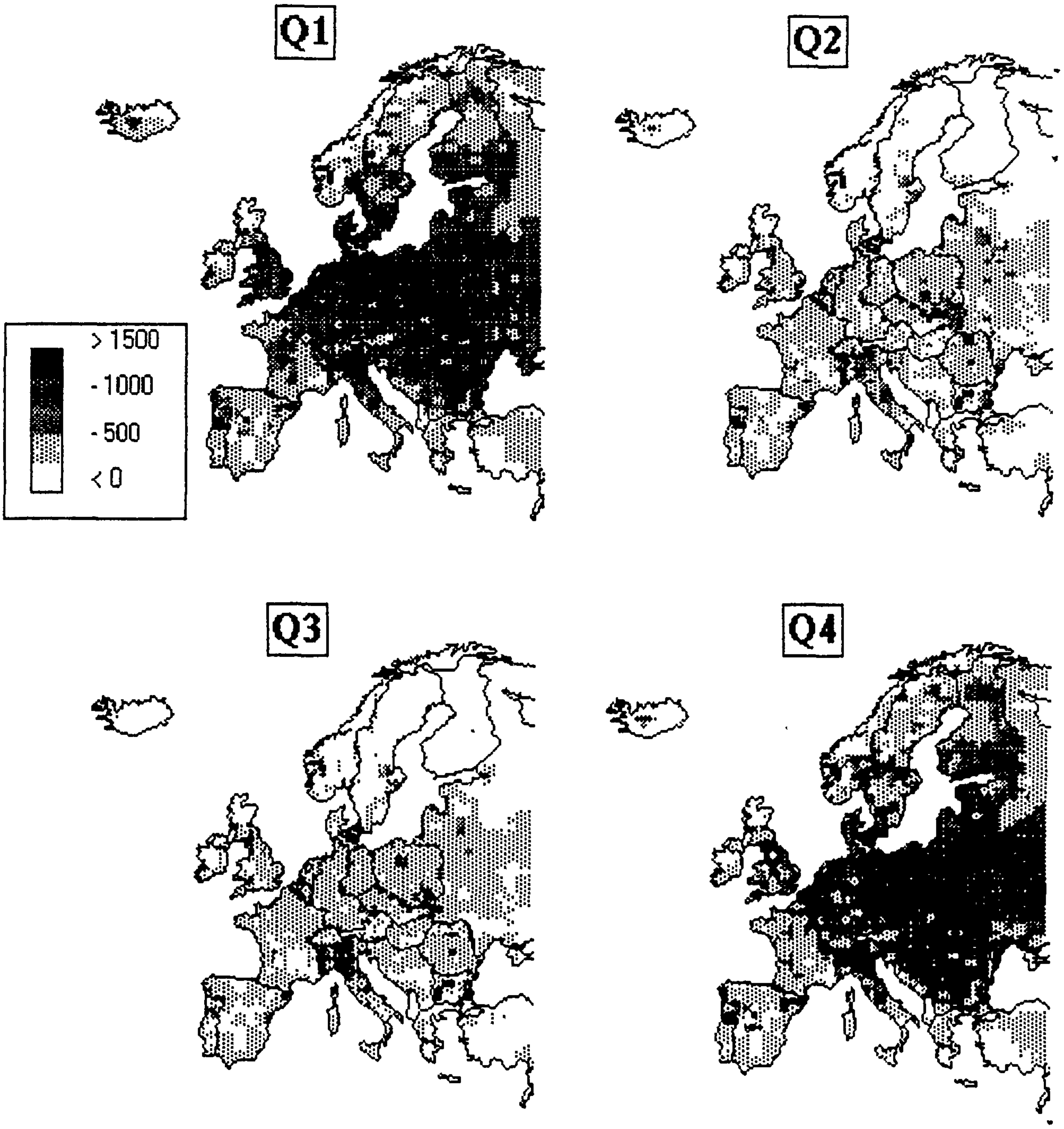

Figure 6. Contour maps of extinction coefficient (10^{-6} m^{-1}) for unfiltered data.

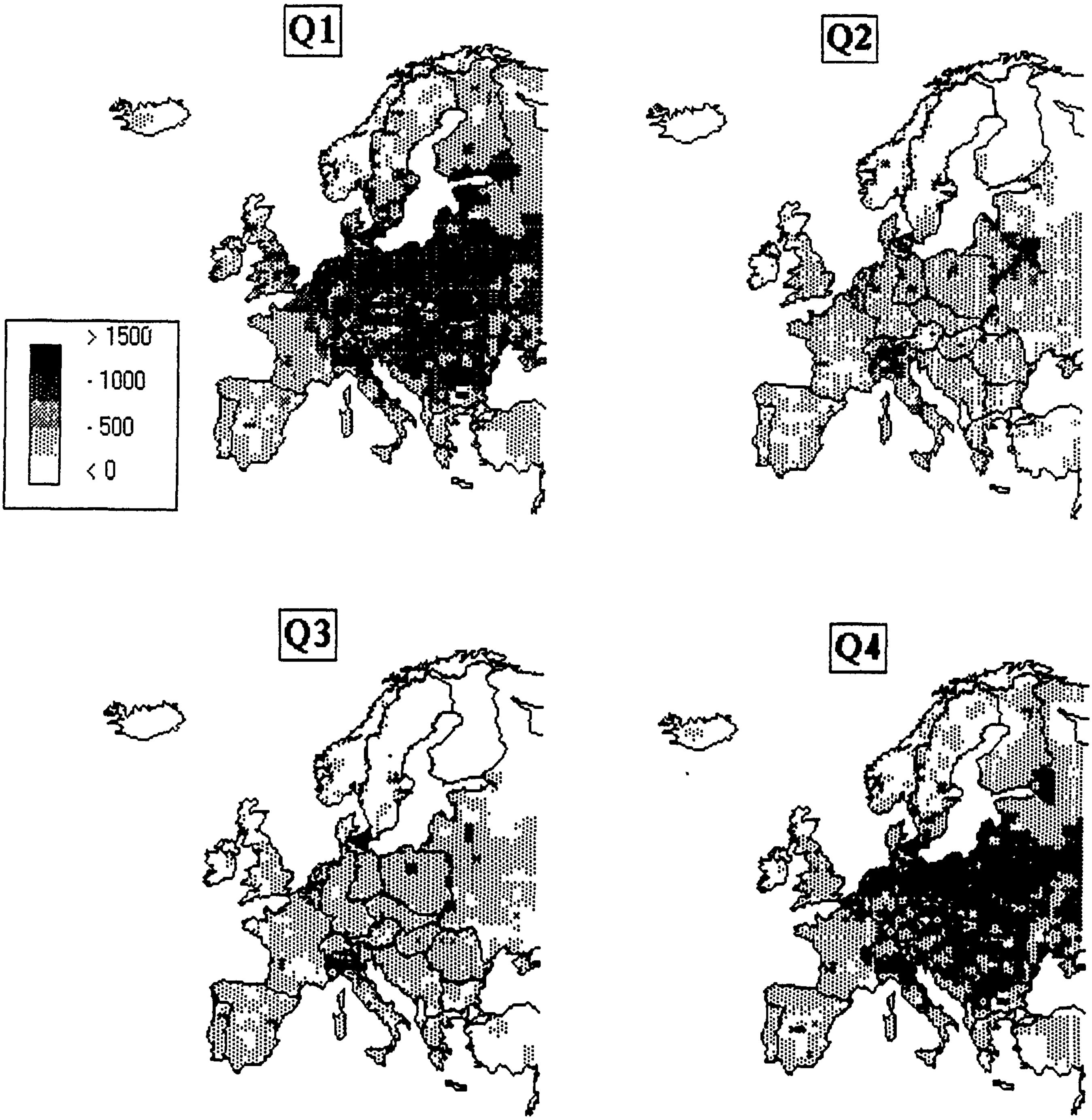

Figure 7. Some contour maps of extinction coefficient (10^{-6} m^{-1}) after imposing the precipitation and humidity filters.

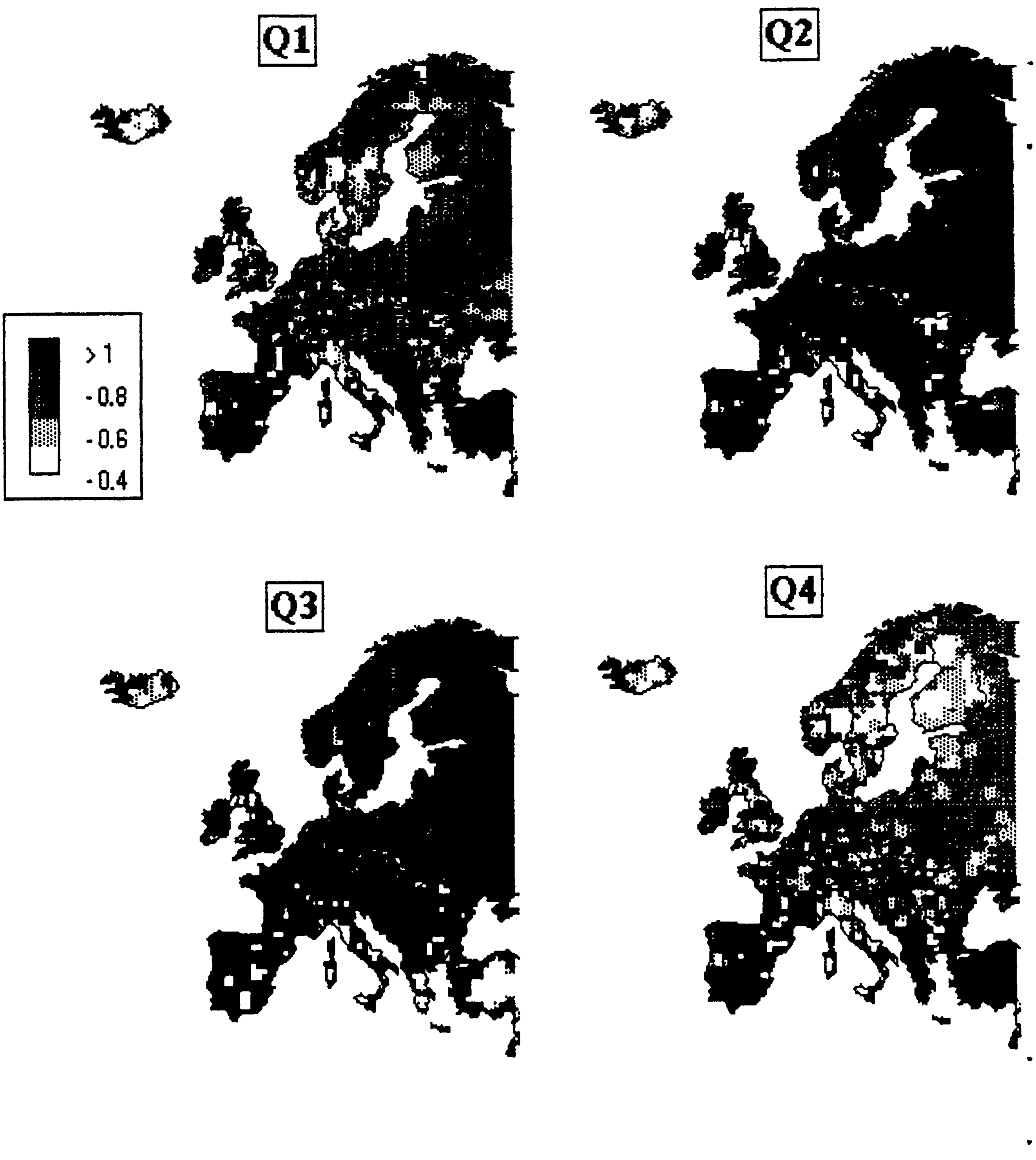

Figure 8. Ratio of precipitation filtered data to raw data.

contains the same maps but after the data passed this filter. Figure 8 shows the ratio of the B_{ext} after passing this filter to raw data. As it can be seen, the precipitation filter did not change the contour of the B_{ext} over Europe for quarters 2 and 3, but reduced the "hot spots" such as those in Bulgaria, Britain, and northern Italy in quarter 3. This filter decreased the extinction coefficient over a much larger area across Europe during quarters 1 and 4. In quarter 1, for example, the extinction coefficient exceeded $1000e^{-6}$ m^{-1} over large areas of Eastern Europe. After the precipitation and humidity filter, most of these high values were eliminated. This seasonality of this filter is best illustrated in Figure 8. This figure shows that quarters 1 and 4 were heavily filtered, while the data in quarters 2 and 3 bypassed this filter except in some isolated spots. Regrettably, during the initial computations on the hourly data the number of discarded data points were not recorded. Consequently, we can not quantify the fraction of hourly values discarded.

3.2 Time Trend Filters

The visibility data base covers the time span of 14 years for most stations (1973-1986). During this time span the 75th percentile shows a specific time pattern for each station. The behavior of such a time trend can be used to detect and eliminate anomalous data points or systematically biased data. An example of systematic effects of visibility threshold was discussed in detail in section 2.

The filters discussed in this section all operate on the time trend of the 75th percentile. They eliminate lone data points, data points that don't vary with time, and outliers in the time chart. The effect of the data filters 2-7 is summarized in Table 2. The top section of the table shows the results of the time trend filters 2, 3, and 4. These filters operate on data points, that is the extinction coefficient for a specific quarter, year, and station. The total number of such data points is about 16000 per quarter. Filters 5, 6, and 7 in the bottom halve of the table eliminate entire stations, not only data points.

Table 2. The effect of the data filters 2 - 7

Data Point Filters

	Q1		Q2		Q3		Q4	
	Points	%	Points	%	Points	%	Points	%
Initial Data Pts.	16709	100	16343	100	15893	100	17087	
Removed by Filter 2	169	1.0	154	1.0	204	1.3	87	0.5
Removed by Filter 3	414	2.5	275	1.7	307	2.0	530	3.1
Removed by Filter 4	3481	21.0	5023	31.0	5073	32.0	3922	23
All Point Filters	4064	24.3	5452	19.0	5584	35.1	4539	27.7

Station Filters

	Q1		Q2		Q3		Q4	
Station Filters								
Stations Before	1578		1574		1569		1585	
Removed by Filter 5	18	1.1	18	1.1	18	1.1	18	1.1
Removed by Filter 6	240	15.2	349	22.2	396	25.2	260	16.4
Removed by Filter 7	57	3.6	39	2.5	34	2.2	54	3.4
All Station Filters	315	20	406	26.8	448	28.5	332	21

Filter 2. **Single point filter.** This filter deletes all points which do not have at least one observation two years before and after it.

Filter 3. **Spike filter.** This filter discards all observations where the B_{ext} is greater than or equal to two times the value of the observations one year before and after it.

Filter 4. **Straight line filter.** This filter discards all observations which have the same value for three or consecutive years

Filter 5. **The Elevation Filter.** This filter eliminates stations in the alpine regions of France, Switzerland, and Austria which are at elevations over about 2000 meters

Filter 6. **Station filter.** This filter eliminates all stations which contain less then three years of data.

Filter 7. **Coefficient of variation filter.** This filter discards all stations with a coefficient of variation greater than 50 percent and less than 1 percent.

3.2.1 Single Point Filter (2)

This filter removes single data points with missing data before and after it. Many stations have sporadic data coverage over the fourteen year period. It is our contention that lone data points, unsupported by neighboring years, are not reliable measures. Consequently the single point filter was implemented to eliminates all data effect of this filter is shown on the time trend charts, Figure 9. As seen in Table 2 the point filter removed about 1% of the quarterly data points.

3.2.2 Spike Filter (3)

This filter detects and discards spikes in the time trends. This filter eliminates data points that are a factor of two larger than their neighboring points. At each point this filter checked the value of the points one year before and after it. If the value of the point being check was greater than twice the value of the other two points, it was deleted. These points were deleted because after examining the data it was assumed that fluctuations of this magnitude was abnormal. It is possible that these fluctuation were caused by observation or recording error. Figure 10 shows the time plot of a station before and after this filter was applied. We recognize that selecting a factor of 2 as the noise filter criteria is rather stringent and subjective. As seen in Table 2 the spike filter removed about 2% of the quarterly data points.

3.2.3 Threshold Filter (4)

The purpose of this filter is to eliminate data points that do not vary over three or more consecutive years. Its purpose is to detect data biased by the visibility threshhold. Figure 11 contains the time charts of a station having a "flat spot", before and after the filter was applied.

In section 2.4 we have illustrated the fact that a consequence of the visibility threshold is a time invariant time series (see Figure 4). This filter is intended to eliminate those readings that may have been influenced by such a threshold.

This filter had a significant effect on the data. It eliminated between 20 and 35 percent of the observations depending on the quarter as seen in Table 2. We recognize that the criteria imposed on this filter were possibly more stringent than necessary. For instance one could have used four or five years of consecutive constant values for data elimination.

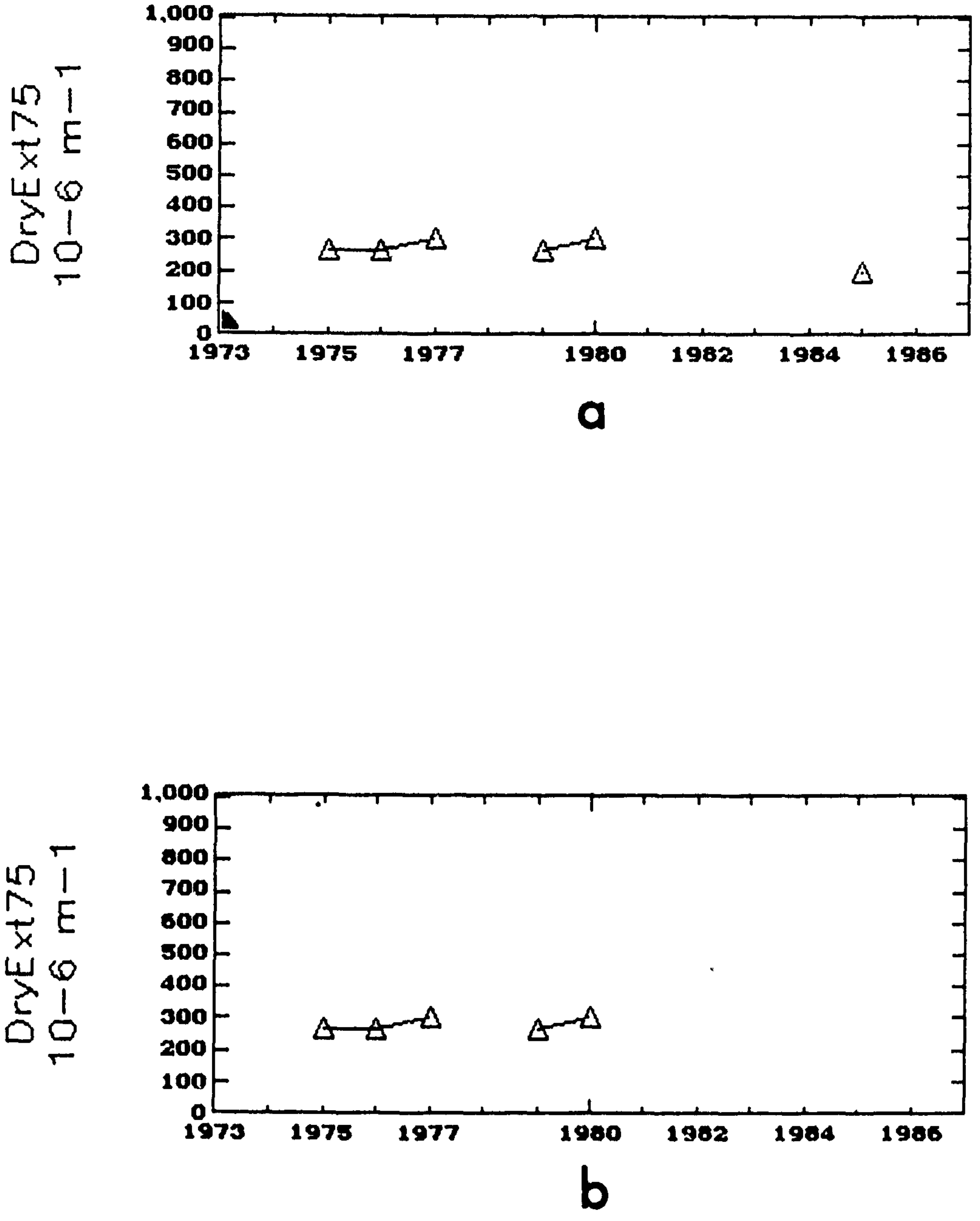

Figure 9. Illustration of the time trend filter a) a station before passing the filter, b) the same station after passing the filter. The value at the year 1985 was deleted since it could not be compared to any values in its vicinity.

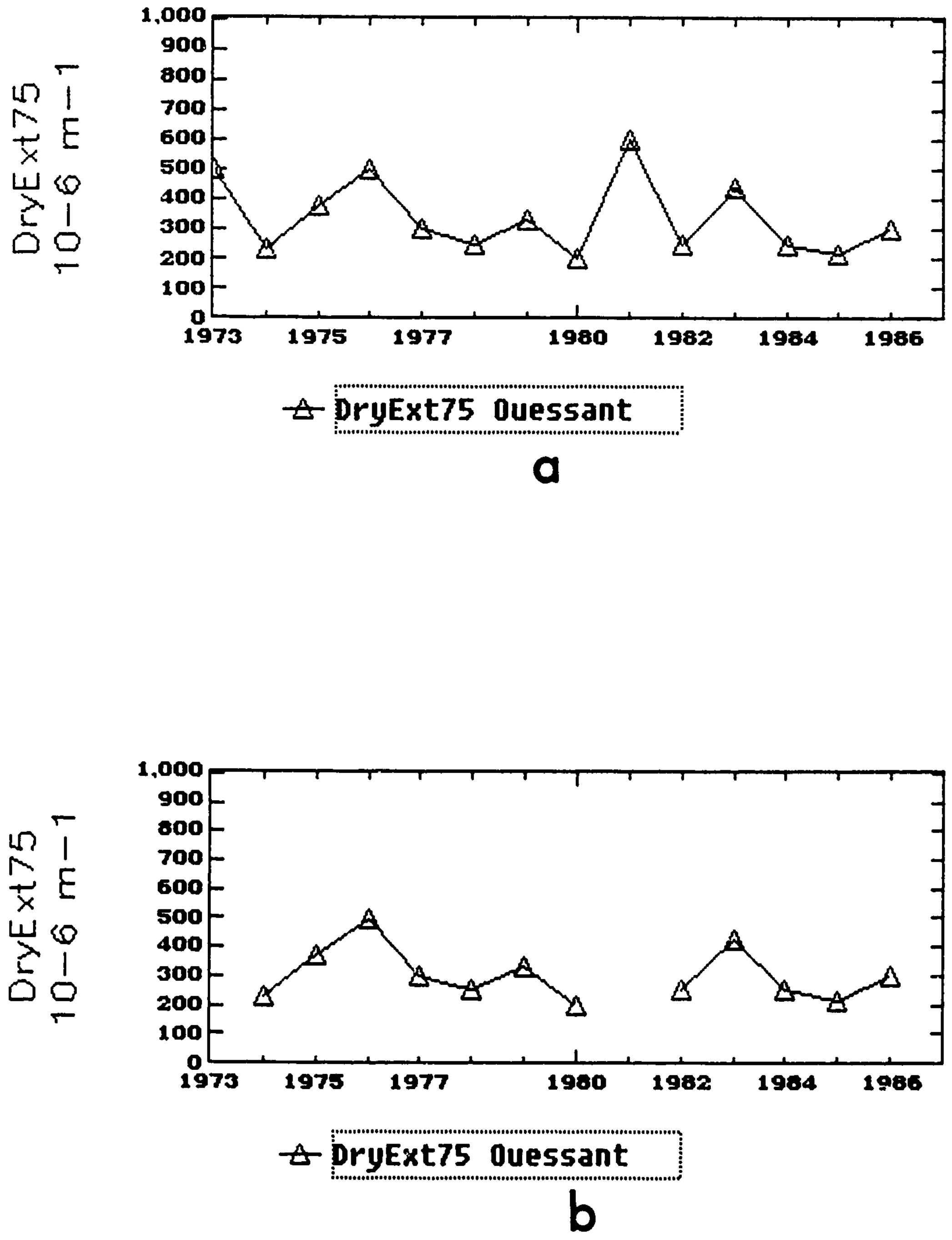

Figure 10. Illustration of the spike filter a) a station before passing the spike filter, b) the station after passing the filter.

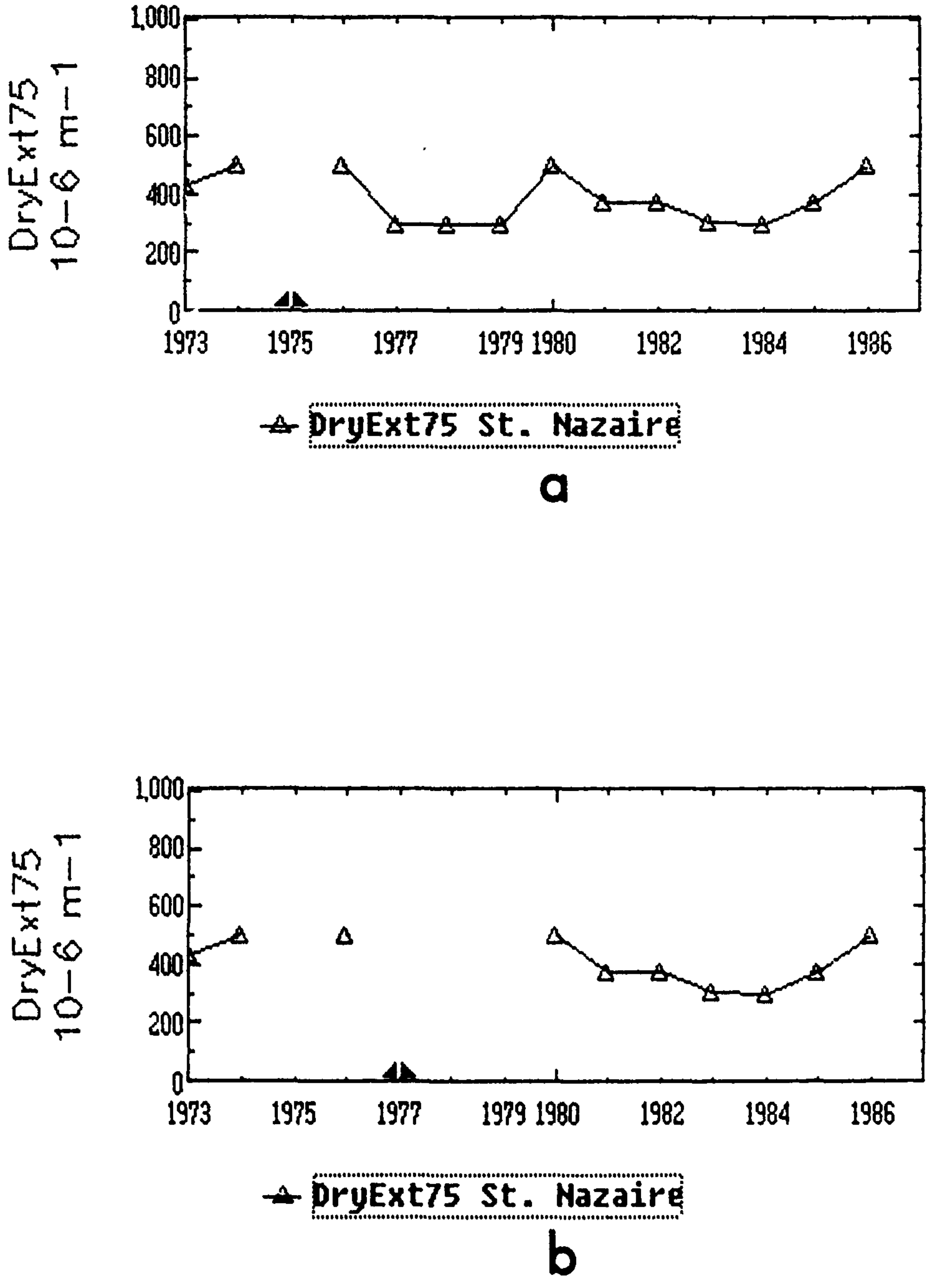

Figure 11. Threshold filter a) a time plot of a station before passing the threshold filter, b) after passing the threshold filter.

3.2.4 Combined Effect of Time Trend Filters

The overall effect of these time trend filters is shown in Figures 12-14 and Table 2. Figure 12 shows the number of years that data existed at each station (maximum of 14) before the time trend filters were applied. On the average over Europe, data were recorded 10 out of 14 years. Notably, less coverage is over Norway and Sweden. Figure 13 shows the number of data points after the filters were applied. After the data passed these filters only central Europe and Finland continued to have high data coverage. Over much of Eastern Europe, the number of data years was below 5. The effect of these filters is best seen in Figure 14 showing the ratio of data years before and after the filters. Over central and western Europe, and Finland, more than 75% of the data have passed these filters. In eastern Europe this percentage was less then 50.

3.3 Station Filters

The following filters apply to stations. That means that all of the values measured at the station were eliminated. These filters utilize the overall characteristics of a station, such as elevation, number of data points, and overall variation of time trends for the stations.

3.3.1 Station Elevation Filter (5)

Observation sites located at elevations over about 2000 meters are generally above the mixing layer where most of the aerosols reside. The extinction coefficients at those sights are significantly lower than neighboring stations at lower elevations. The station elevation filter was imposed to eliminate such high altitude stations on the grounds that they are not representative.

The stations were singled out by comparing the average B_{ext} over all years for each station with that of the surrounding stations in the map view of Voyager. Any station whose average B_{ext} was about a factor of three less than the surrounding stations, were deleted from the data base. This filter eliminated 18 out of about 1600 stations in Europe. Figure 15 shows two of the 18 stations (circled), Jungfraujoch in Switzerland and Hahnenkamm Mountain in Austria, which were discarded using this method. This filter was only applied to the alpine regions of France, Switzerland, and

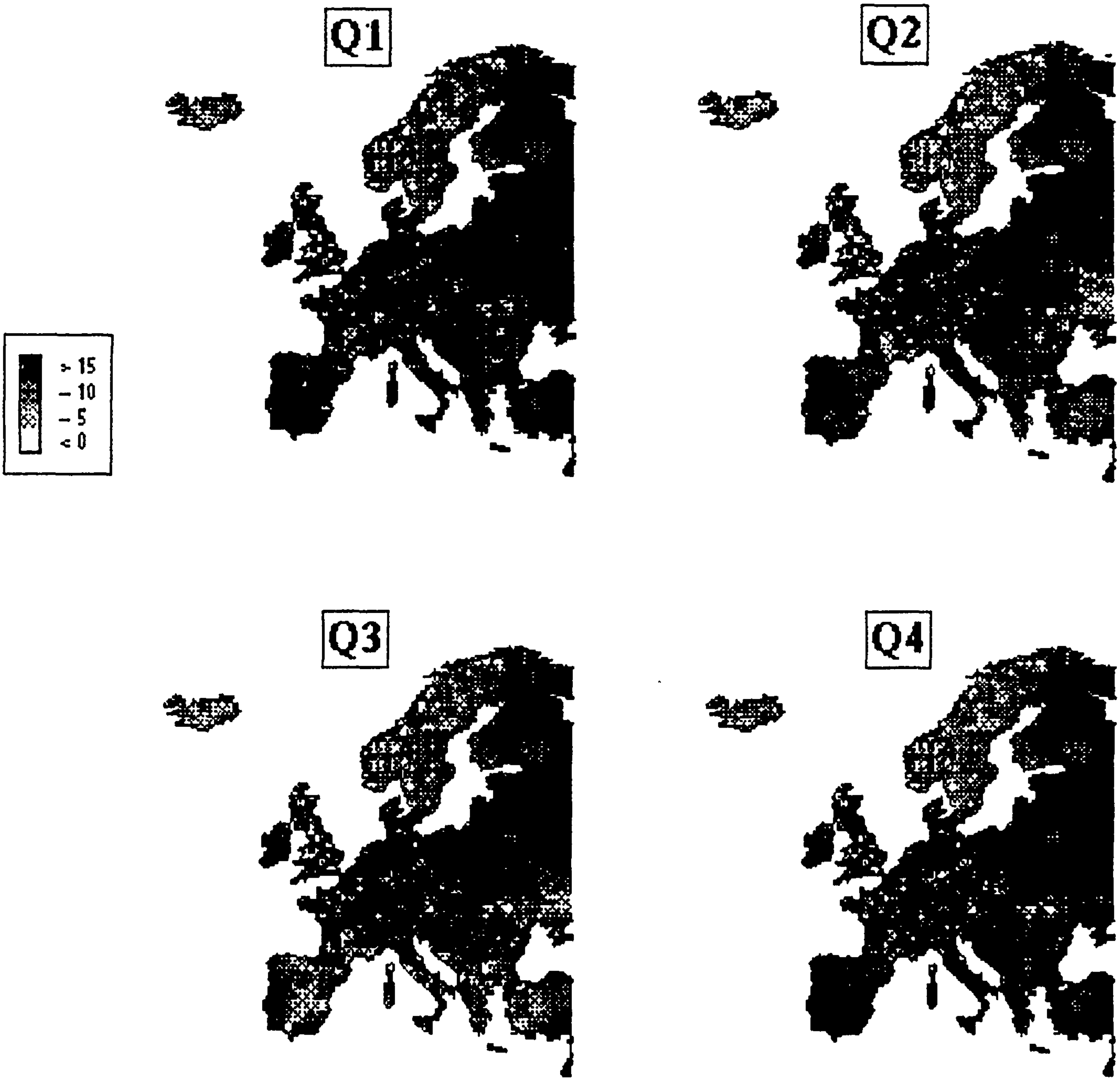

Figure 12. Maps for each quarter of the number of years that data existed at each station before passing time trend filters.

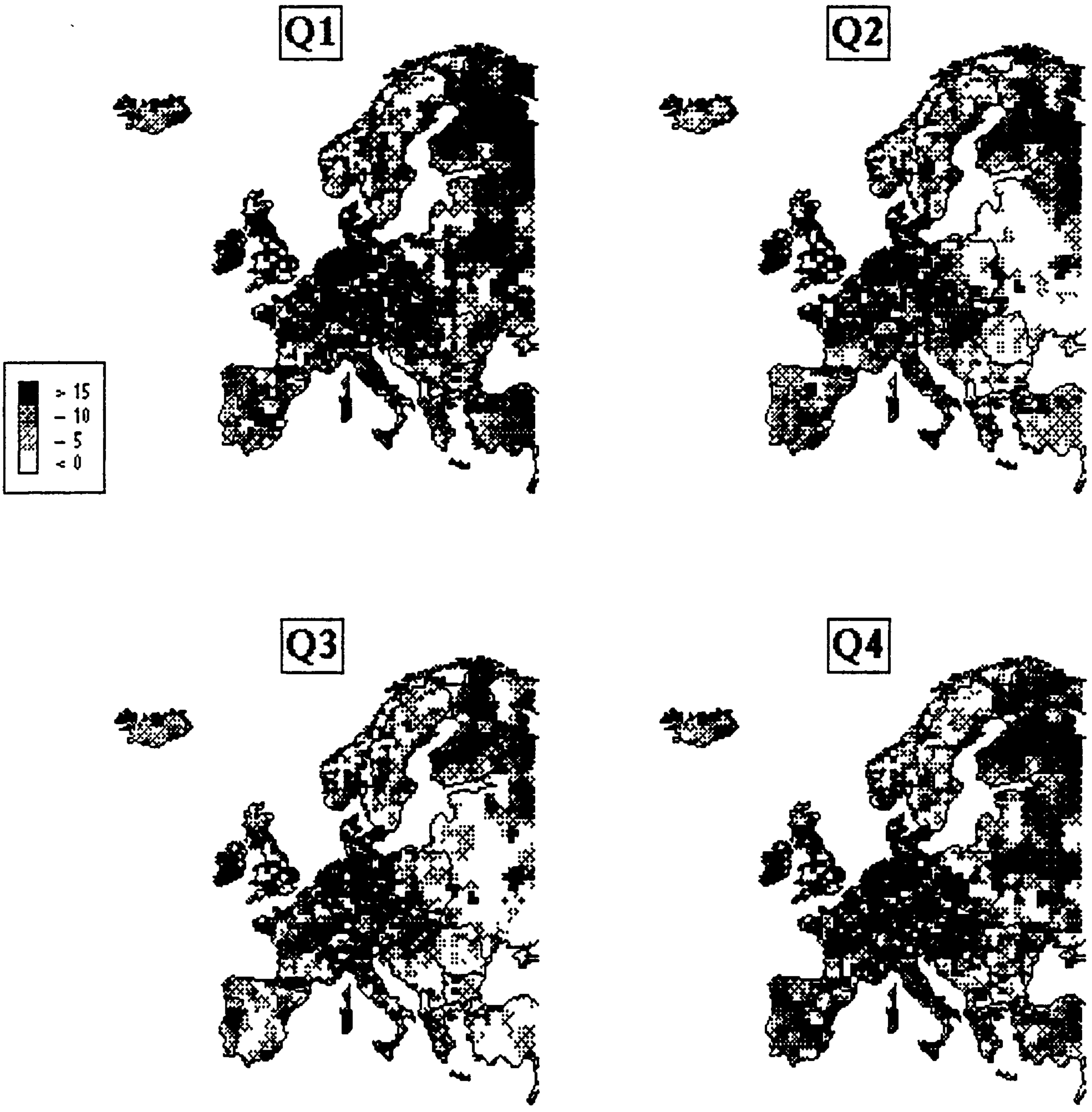

Figure 13. Maps of each quarter of the number of points that data existed at each station after passing time trend filters.

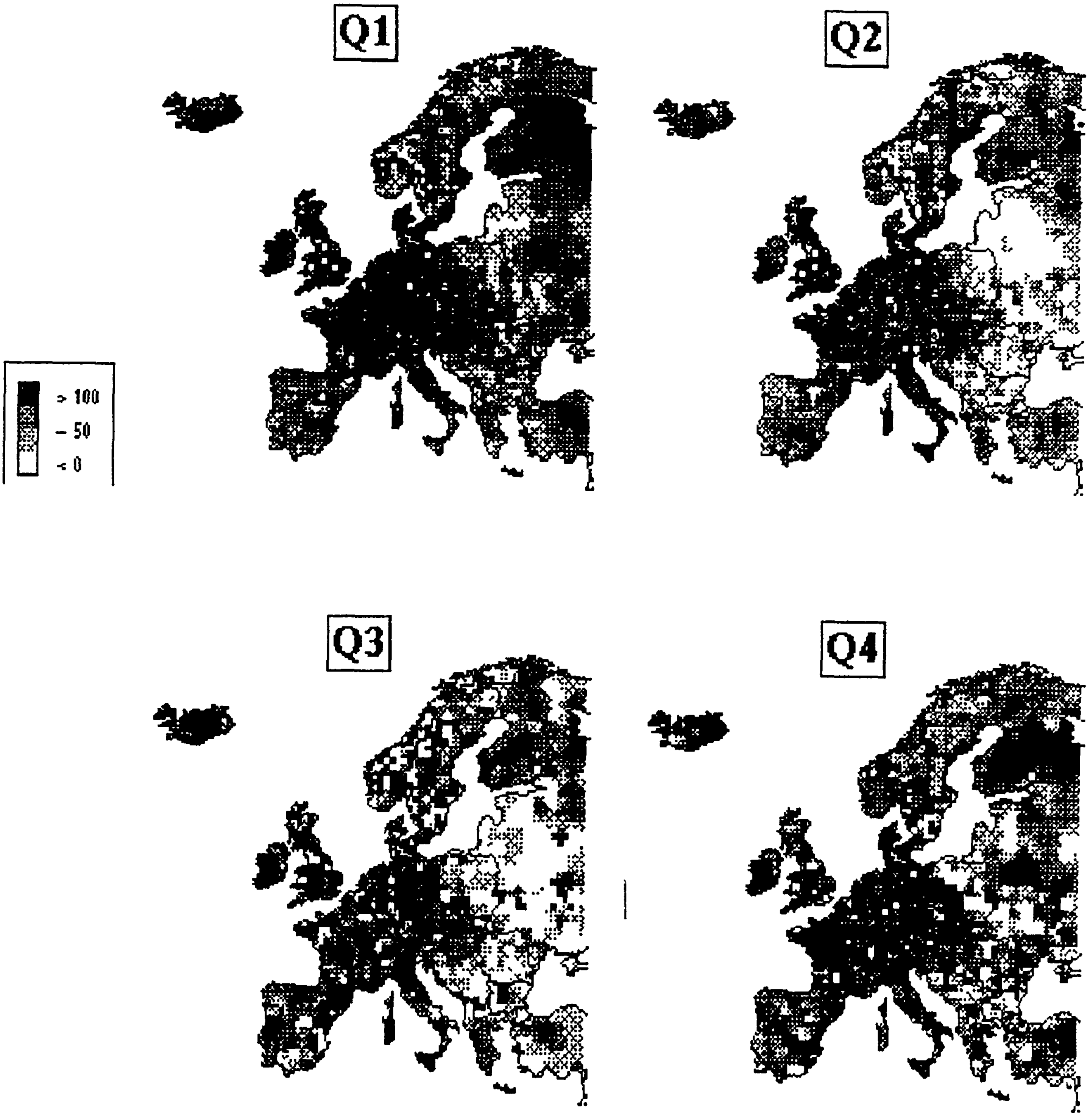

Figure 14. Ratio of the number of points after to the number of points before passing time trend filters.

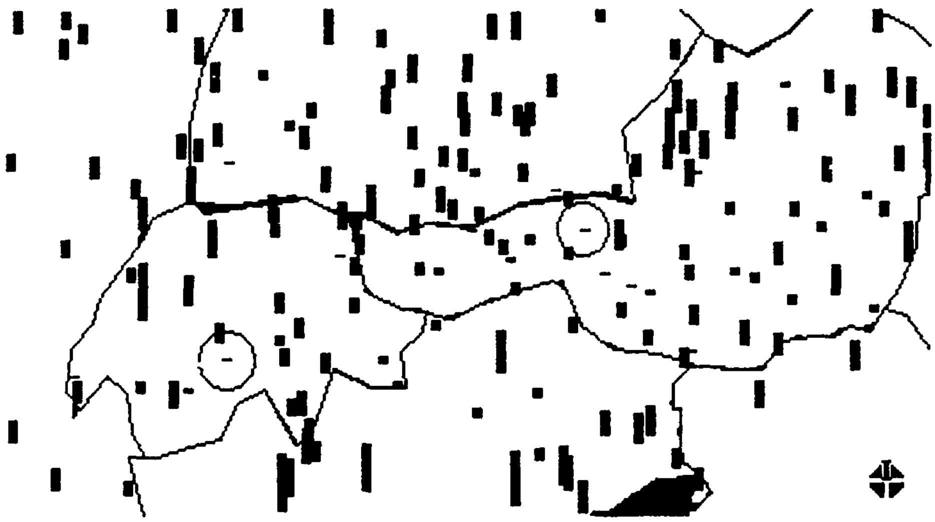

Figure 15. Two stations, Jangfraujoch Switzerland and Hahnenkamm Mountain, Austria, that were discarded by the elevation filter.

Austria. As seen from Table 2, the station elevation filter eliminated an insignificant fraction of the European stations.

3.3.2 Insufficient Data Filter (6)

This filter eliminates stations that have less than three data points over the fourteen year period. It is argued that one or two data points do not provide an adequate characterization of a station. It should be noted that this is one of the last filters applied to the data set. As seen in Table 2, a significant fraction, about 30% of the stations were removed by this filter.

3.3.3 Noise Filter (7)

This filter eliminates those stations that show either excessive variation or too little variation over the fourteen year period. A station is declared to have excessive variation if the standard deviation of the existing data in the fourteen year observation period divided by the mean (coefficient of variation) is greater than 50%. These stations are eliminated on the grounds that they are too noisy. Stations for which the coefficient of variation over the fourteen years is less than one percent are also eliminated. This condition is intended to catch any stations that have passed the threshold filter (4). About 3% of the 1600 stations were eliminated by this filter.

Figure 16 shows the coefficient of variation of the 75th percentile for all four quarters. The highest yearly variation is noted for Scandinavia where it exceeds 40 percent for all seasons. During quarters one and four south central Europe, including Yugoslavia, Romania, and Bulgaria also exhibit similarly high variations. The coefficient of variation over western and southern Europe is between 20 and 40 percent.

3.3.4 Combined Effect of the Station Filters

The effect the three station filters (5, 6, and 7) had on the data base is shown in Figures 17 and 18. Figure 17 shows all station containing valid data. Figure 18 shows which of these stations were discarded for each quarter by the three station filters. As can be seen, a high percentage of station were removed from the coast of Norway, the Alps region and southeastern Europe. Germany and Southern Italy, two areas with high station concentrations, had very few stations discarded.

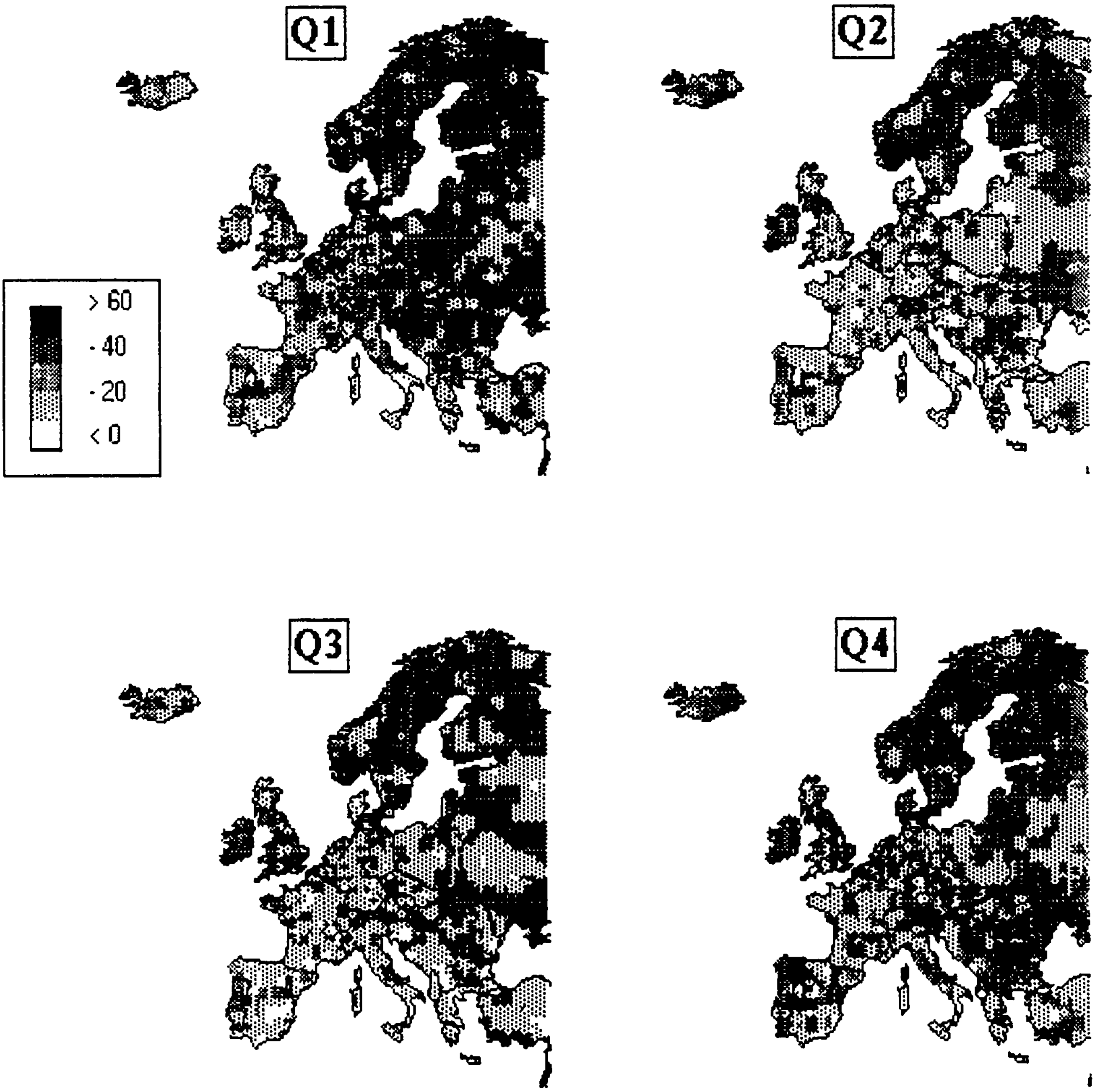

Figure 16. The spatial pattern of the coefficient of variation of the full filtered 75th percentile B_{ext} data.

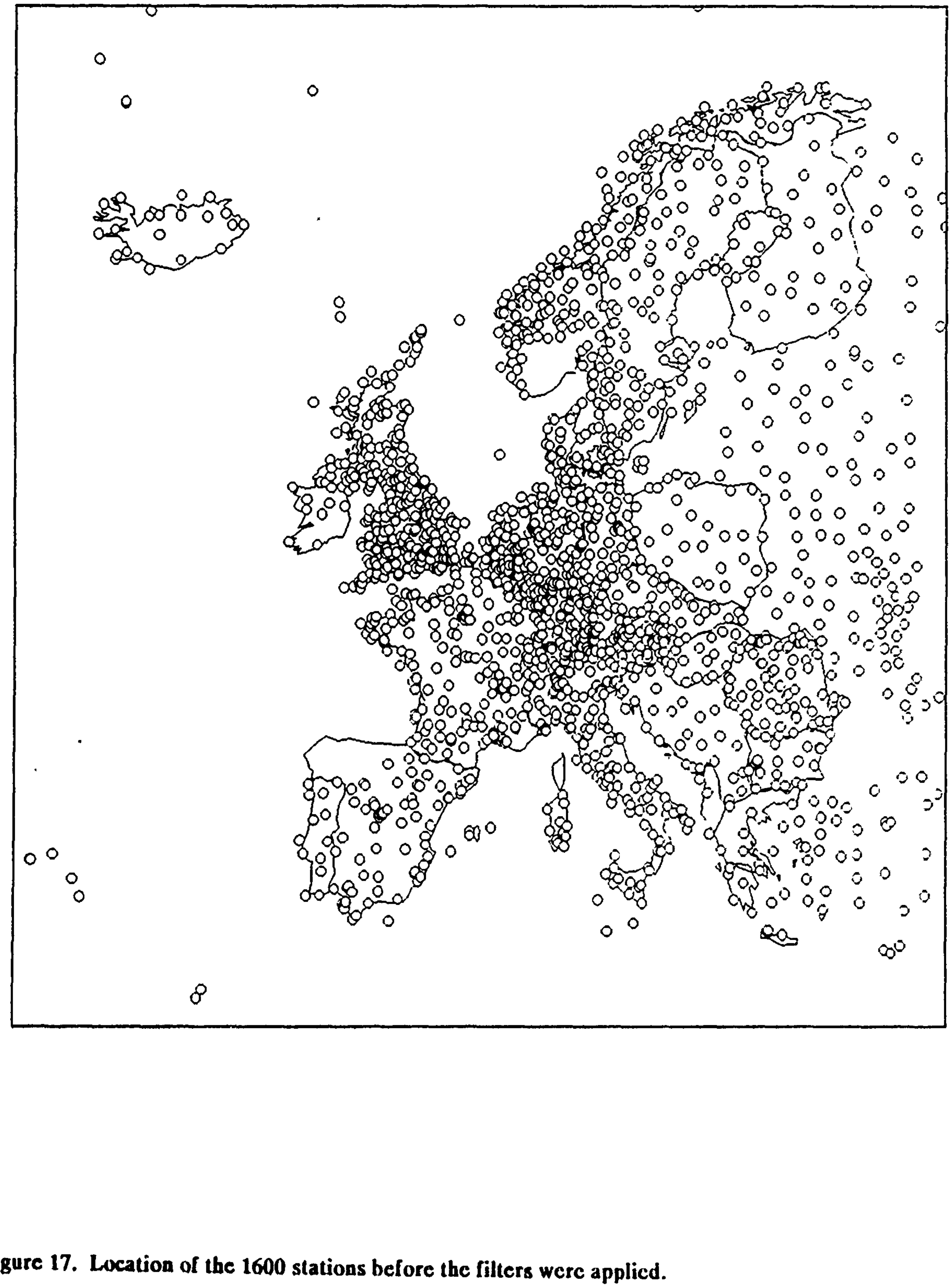

Figure 17. Location of the 1600 stations before the filters were applied.

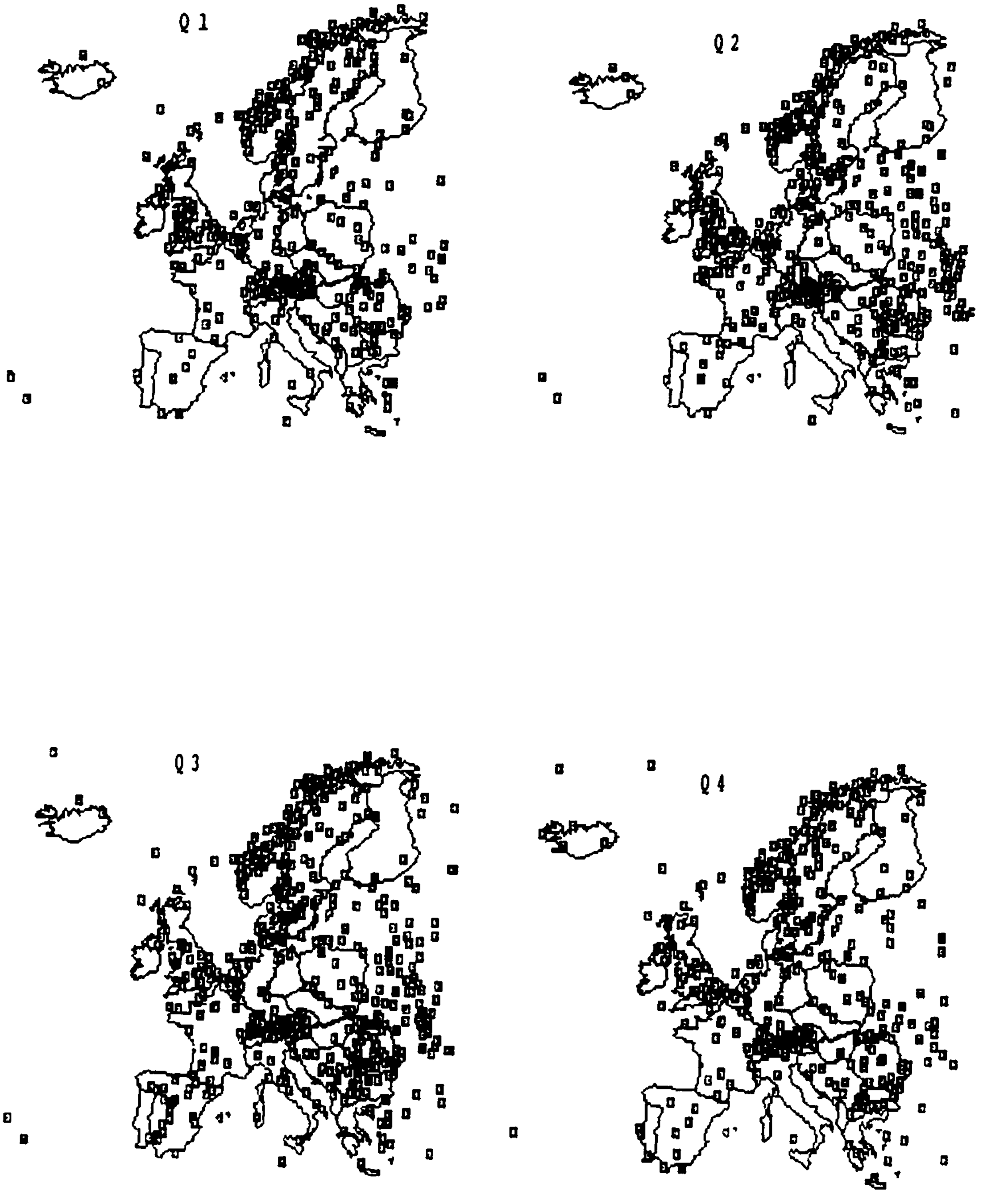

Figure 18. Location of the stations containing data that were discarded by the station filters.

4. SUMMARY AND DISCUSSION

The purpose of the report was to present the methodology for cleaning up the meteorological visibility data from undesirable and erroneous data. Data filters were devised and imposed on the European synoptic visibility data set. The data set consisted of fourteen years of meteorological data (1973-1986) for about 1600 station in Europe. This European data set was extracted from the DATSAV global weather database maintained by the U.S. Air Force, ETAC, Scott Air Force Base..

The raw meteorological data set consisted of over 1000 magnetic tapes containing about 30 gigabytes of data. The first step in the data processing involved compacting the data set into a binary form, which reduced the data size to a more manageable 3 gigabytes. Next, from the daily visibility data, the cumulative distribution functions for extinction coefficient were computed. Most of the subsequent data filtering was performed using the distribution functions and the Voyager data exploration software.

The data cleaning filters fell into three categories: 1. precipitation and humidity filters, 2. filters based on the year to year fluctuation of extinction coefficient, and 3. filters that eliminated entire stations. The precipitation and humidity filter was imposed on the hourly data, while the remaining filters operated on the distribution functions.

The main cause of poor data is identified to be the visibility threshold, that is the maximum distance reported for a given station. The visibility threshold was detected using the shape and time trend of the distribution functions. A visibility threshold truncates the distribution function and also causes the lower percentiles to be invariant with time. It was determined that the 75th percentile of B_{ext} is a robust measure of the extinction coefficient, relatively uninfluenced by the visibility threshold.

This summary section presents the spatial pattern of extinction coefficient for each quarter following the application of all data filters. It also discusses the differences between the extinction coefficients of raw and filtered data.

The quarterly maps, Figure 19, represent the extinction coefficients after the application of all of the filters. It shows that the average B_{ext} is highest over Europe during quarter one (January, February, and March), and the lowest during quarter two

(April, May, and June). Quarter four has a similar pattern to quarter one and quarter three resembles quarter two. In this sense the extinction coefficient over Europe could be lumped into a cold season (October-March), and the warm season (April-September).

After application of all of the data quality and metrological filters the highest extinction coefficient is observed over Northern Italy. The Po River Valley is an industrial hot spot, and the cause of the high extinction coefficient there is undoubtedly man induced air pollution. Other areas of high extinction coefficient cover the "coal belt of Europe" stretching from southeastern Great Britain through Germany, and Poland. Another area of high extinction coefficient covers Romania and Bulgaria.

The corresponding maps of raw data prior to the application of the data to filters was presented in Figure 6. The comparison of the sets of maps of Figures 19 and 6 shows the changes caused by the precipitation and data quality filters. It is of considerable interest to examine the nature of the influence of these filters including spatial pattern and magnitude. Inspection of the two sets of maps reveals similar spatial pattern of the European extinction coefficient for the filtered and unfiltered data. However, the absolute magnitude of the extinction coefficient is about 20 to 50% lower for the filtered data. The spatial pattern also reveals that this 20 to 50% difference is quite consistent over all geographic areas and all quarters.

It is comforting to observe that the application of data filters did not change the qualitative pattern of extinction coefficient over Europe, they merely influenced the overall magnitudes. This observation somewhat relieves the pressure of justification for the seven data filters. Recognizing the subjective manner in which several of the data filters were chosen, their full justification could be very demanding.

Future reports in this series will focus on the detailed presentation of the resulting "clean" extinction coefficient database for Europe and North America. The apportionment of the extinction coefficient into different aerosol types will also be presented. With such documentation, the present aerosol database will be suitable for application in radiative transmission models, atmospheric climate models and other studies.

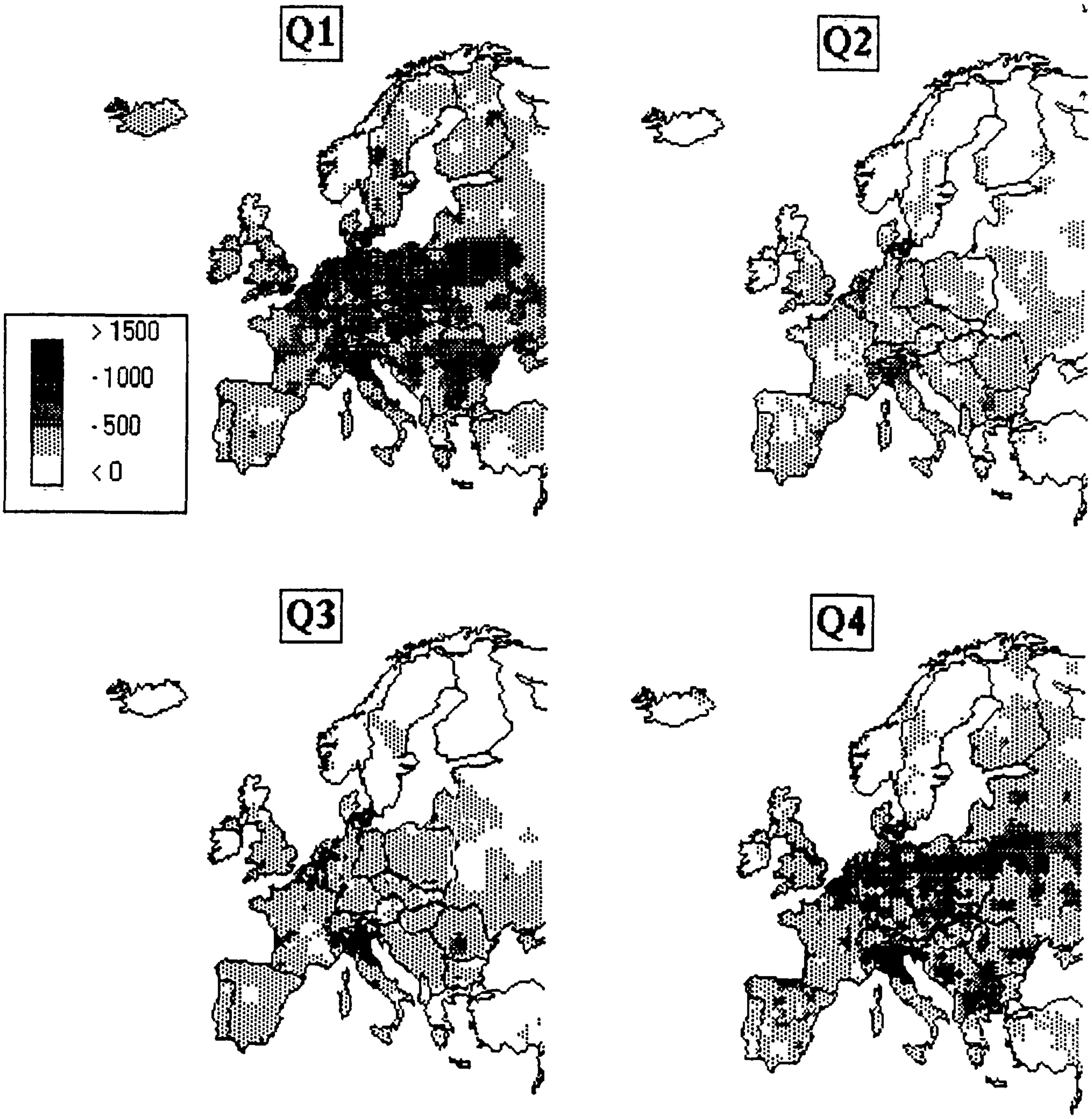

Figure 19. Maps of filtered extinction coefficients for Europe.

APPENDIX A: CONTOURING OF SPATIAL DATA

Meteorological variables are measured at specific stations dispersed randomly over a geographic region. The presentation of such spatial data can be aided significantly by spatial extrapolation or contouring. Contoured data represents smoothed patterns and illustrates the high and low regions as well as the spatial gradients. Such pattern can not be discerned from presentation of individual data points.

Obtaining a contour map from measurements at random locations requires a spatial extrapolation procedure. In this project we used the contour program running on an IBM PC to perform the contouring. The Contourer is an interactive Windows program that produces shaded contour plots from a set of random data points. It first projects the data onto a uniform grid and then draws the contours. Data for the array of points can be supplied as ASCII files from Voyager's map view, from spreadsheets, or from other programs.

The contour program requires an ASCII table containing at least the latitude, longitude, and the parameter value at each location as inputs. The contourer uses map data files prepared by the Voyager map compiler.

A.1 The Extrapolation Algorithm

Contouring of data is part art, part science. It is a spatial extrapolation process by which the value at an unknown location is extrapolated, based on the measured values at neighboring locations. The spatial extrapolation can be done in many different ways. In the Contourer, the extrapolation is accomplished through the construction of a grid under the spatial data domain. The values are calculated for every grid point and plotted by a shading algorithm. Higher values at the grid point are assigned darker shades.

The value at a given grid point is obtained using a spatial filter: the grid value, g_j, is weighed average of surrounding data point, c_i, where the weighing factors are w_{ij}

$$g_j = \frac{\Sigma_i w_{ij} c_i}{\Sigma_i w_{ij}}$$

The weighing factors are power law functions of distance from grid to the neighboring site R_{ij}, such that

$$w_{ij} = \frac{1}{R^n_{ij}}$$

where n is a user selectable exponent, ranging between 0 and 3. Larger n results in less spatial smoothing, more texture of the grid values.

The extrapolation is constrained by a user selectable radius of influence, R. Stations outside that radius are not considered in the weighing. Its purpose is to prevent extrapolation beyond "reasonable" distances from measured data.

The maximum number of stations used in the weighing is also user selectable. Suppose that the number of stations within the search radius is 20, but you set the maximum number of stations to be 6. This means that only the closest 6 stations are used. This procedure allows varying the smoothing texture with station density.

The minimum number of locations for a valid grid point can also be set. If this value is set to 2, for example, then every valid grid point has to have at least two stations within its radius of influence.

The contour maps can be printed by cutting the bitmap from the Contourer and posting it into Windows program Paint which has printing facilities.

REFERENCES

1. USAFETAC DATSAV Database Handbook (1977) Unites States Air Force, Air Weather Service, USAF Environmental Technical Application Center, Scott Air Force Base, IL 62225, USAFETAC-TN-77-2.

2 Kneizys F.X., Shettle E., Gallery W.O., Chetwynd, Jr. J.H., Abreau, L.W., Selby, J.E.A., Clough, S.A., and Fenn, R.W. (1983) Atmospheric Transmittance/Radiance: Computer Code LOWTRAN 6, Air Force Geophysics Laboratory, Hanscom AFB, MA 01731, AFGL-TR-83-0187, ADA137786.

3. Middleton W.E.K.(1952) Vision Through the Atmosphere. University of Toronto Press, Toronto, Canada.

4. Husar, R.B., Patterson, D.E., Holloway, J.M., Wilson, W.E., Ellestad, T.E. (1979) Trends of eatern U.S. haziness since 1948. 4th Symposium on Turbulence Diffusion and Air Pollution. Reno, NV.

5. Malm, W.C., Walther, E.G., O'Dell, K, Kleine, M (1981) Visibility in the southwestern United States from summer 1978 to spring 1979. Atmospheric Environment, 10/11, 2031-2042.

6 Morales, C., Husar, R.B. and El-Ghazzaway, O. (1986) Use of visibility observations for the investigation of hazy air masses. Department of Meteorology, University of Stockholm, Stockholm, Sweden, ISSN0280-445X.

7. Husar, R.B., Oberman T., and Hutchins, E.A. (1990) Environmental Informatics: Implementation Through the Voyager Data Exploration Software. Air and Waste Management Association proceedings. Pittsburgh, June 25-29.